

# robots.txt - Robots kontrollieren

## Allgemeines zur robots.txt

Es gibt Defacto-Standards im Internet, die einfach gewachsen sind, ohne es je zu einer [RFC](#) gebracht haben. Dazu gehört auch der Status, den die Datei *robots.txt* im Web hat. In einer Datei dieses Namens können Betreiber von Web-Projekten angeben, welcher Such-Robot welche Projektverzeichnisse auslesen darf und welcher was nicht lesen darf. Die Datei enthält also Anweisungen für Robots von Suchmaschinen. Die überwiegende Mehrheit der Robots moderner Suchmaschinen berücksichtigen das Vorhandensein einer *robots.txt*, lesen sie aus und befolgen die Anweisungen.

Zwar lässt sich auch in einzelnen HTML-Dateien mit Hilfe eines Meta-Tags [für Suchprogramme das Auslesen erlauben bzw. verbieten](#). Doch das betrifft nur die jeweilige HTML-Datei und maximal alle weiteren, durch Verweise erreichbaren Dateien. In einer zentralen *robots.txt* können Sie dagegen unabhängig von der Datei- und Verweisstruktur Ihres Web-Projekts festlegen, welche Verzeichnisse und Verzeichnisbäume ausgelesen werden dürfen, und welche nicht.

## Speicherort und Aufbau einer robots.txt

Die *robots.txt* muss unter diesem Namen (alle Buchstaben klein geschrieben) im Wurzelverzeichnis der Web-Dateien der Domain abgelegt werden. Wenn Sie also einen Domain-Namen *mein-name.de* haben, dann muss die *robots.txt* in dem Verzeichnis abgelegt werden, in dem auch die oberste Einstiegsdatei von *www.mein-name.de* liegt. Der URI wäre also *http://www.mein-name.de/robots.txt*. Nur so kann sie von Suchmaschinen-Robots, die das Projekt aufsuchen, gefunden werden. Das bedeutet, dass Sie die Technik der *robots.txt* nur nutzen können, wenn Sie eine eigene Domain haben, nicht aber bei Webspace-Angeboten, wo Sie lediglich ein Homepage-Verzeichnis auf einem Server erhalten, ohne an das Wurzelverzeichnis der Domain zu kommen.

Die *robots.txt* ist eine reine Textdatei und muss mit einem Texteditor bearbeitet werden.

## Beispiel: robots.txt:

```
# robots.txt zu http://www.mein-name.de/

User-agent: UniversalRobot/1.0
User-agent: mein-Robot
Disallow: /quellen/dtd/

User-agent: *
Disallow: /unsinn/
Disallow: /temp/
Disallow: /newsticker.shtml
```

## Erläuterung

Die erste Zeile ist lediglich eine Kommentarzeile. Kommentarzeilen werden durch ein Gatterzeichen # eingeleitet.

Ansonsten besteht eine *robots.txt* aus Datensätzen (*records*). Im obigen Beispiel sind zwei solcher Datensätze notiert. Ein Datensatz besteht grundsätzlich aus zwei Teilen. Im ersten Teil wird angegeben, für welche Robots (User-agent) die nachfolgenden Anweisungen gelten.

Im zweiten Teil werden die Anweisungen selbst notiert. Die Anweisungen bestehen darin, den zuvor bestimmten Robots etwas zu verbieten (Disallow).

Jede Zeile eines Datensatzes beginnt mit einem der zwei erlaubten Schlüsselwörter User-agent oder Disallow. Dahinter folgt, durch ein Doppelpunkt und Leerzeichen getrennt, die zugehörige Angabe. Zwischen den Datensätzen wird eine Leerzeile notiert. Innerhalb eines Datensatzes muss zunächst mindestens eine Zeile mit User-agent: beginnen. Dahinter ist immer nur eine Angabe möglich. Wenn Sie mehr als einen bestimmten Robot ansprechen möchten, müssen Sie mehrere Zeilen untereinander notieren, die mit User-agent: beginnen - so wie im ersten Datensatz des obigen Beispiels. Unterhalb der Zeilen, die mit User-agent: beginnen, werden die Zeilen notiert, die mit Disallow beginnen. Die Angaben dazu werden dann von den Robots beachtet, die im gleichen Datensatz mit User-agent spezifiziert wurden.

Bei User-agent: ist entweder die Angabe \* (Sternzeichen) erlaubt, was so viel bedeutet wie "alle Robots", oder der Name eines bestimmten Robots. Diesen Namen müssen Sie allerdings kennen. Wenn Sie \* angeben, sollte in dem gleichen Datensatz keine weitere Angabe zu User-agent folgen, da dies dem Platzhalter für "alle" widersprechen würde.

Hinter jeder Zeile, die mit Disallow: beginnt, können Sie jeweils eine Pfadangabe notieren. Die Robots werden diese Pfade auf Ihrer Seite dann nicht indizieren. Es besteht übrigens keine Möglichkeit, Verzeichnisse explizit für die Indizierung zu erlauben. Ein Schlüsselwort Allow: wurde niemals definiert.

Bei den Angaben zu Disallow: können Sie Verzeichnispfade und einzelne Dateien mit Pfadangabe angeben. Wildcards wie \* oder \*.\* sind dabei **nicht** erlaubt. Achten Sie darauf, bei Verzeichnispfaden einen abschließenden Schrägstrich / zu notieren. Wenn Sie nämlich beispielsweise /index notieren, wäre auch die Datei /index.html betroffen, und nicht nur das Unterverzeichnis /index/.

Im ersten der obigen Beispiel-Datensätze wird angenommen, dass ein superschlauer Robot namens UniversalRobot/1.0 sowie ein selbstgestrickter Robot namens mein-Robot Daten aus dem Verzeichnis /quellen/dtd/ in ihre Suchmaschinen einspeisen. Das ist aber nicht erwünscht, weil in diesem Verzeichnis beispielsweise DTDs für XML-Dateien abgespeichert sind. Deshalb wird speziell diesen beiden Robots der Zugriff auf dieses Verzeichnis (und alle Unterverzeichnisse davon) verboten.

Im zweiten Datensatz wird allen Robots verboten, die beiden Unterverzeichnisse /unsinn/ und /temp/ auszulesen. Die Verzeichnisnamen sprechen ja für sich, und es ist leicht ersichtlich, wozu dieses Verbot gut ist: nämlich um überflüssigen Datenmüll in den großen Suchmaschinen zu vermeiden. Ferner wird der Zugriff auf die Datei newsticker.shtml verboten. Der Grund könnte sein, dass diese Datei von einem über [Server Side Includes](#) eingebundenen CGI-Script laufend aktuelle Daten erhält, weswegen es keinen Sinn hat, diese Datei in Suchmaschinen aufzunehmen.

## Beachten Sie:

Mit der folgenden Syntax erlauben Sie keinem Suchmaschinen-Robot, auch nur irgendetwas von ihren Daten auszulesen:


```
User-agent: *  
Disallow: /
```

Mit / bestimmen Sie "alle Daten dieses Verzeichnisses und aller Unterverzeichnisse".

Mit der folgenden Syntax nehmen Sie einen bestimmten Robot namens mein-Robot von allen anderen Verboten aus:

```
User-agent: mein-Robot  
Disallow:
```

Durch eine fehlende Angabe hinter Disallow: wird alles erlaubt!

Web-Browser ignorieren die *robots.txt*. Es ist also nicht möglich, damit Daten vor Anwendern zu schützen. Lesen Sie zu diesem Zweck den Abschnitt  [.htaccess - Server-Reaktionen kontrollieren](#)

Aus demselben Grund besteht auch keinerlei Garantie, dass sich Suchmaschinen an die Verbote in *robots.txt* halten. Ordentlich programmierte Robots beachten die Datei, aber Robots, die mit bösen Absichten das Web durchsuchen, halten sich vermutlich nicht daran. Wenn Sie Informationen sicher vor allen Augen verstecken wollen, richten Sie z.B. einen Passwortschutz ein.